An anecdote about improper capitalization

devblogs.microsoft.com/oldnewthing/20031105-00

November 5, 2003

Raymond Chen

I've already discussed some of the strange consequences of case-sensitive comparisons.

Joe Beda mentioned the Internet Explorer capitalization bug that transformed somebody's name into a dead body. Allow me to elaborate. You might learn something.

This bug occurred because Internet Explorer tried to capitalize the characters in the name "Yamada" but was not mindful of the character-combining rules of the double-byte 932 character set used for Japanese. In this character set, a single glyph can be represented either by one or two bytes. The Roman character "A" is represented by the single byte 0x41. On the other hand, the characters "の" is represented by the two bytes 0x82 0xCC. (You will need to have Japanese fonts installed to see the "no" character properly.)

When you parse a Japanese string in this character set, you need to maintain state. If you see a byte that is marked as a "DBCS lead byte", then it and the byte following must be treated as a single unit. There is no relationship between the character represented by 0xE8 0x41 (錢) and 0xE8 0x61 (鐶) even though the second bytes happen to be related when taken on their own (0x41 = "A" and 0x61 = "a").

Internet Explorer forgot this rule and merely inspected and capitalized each byte independently. So when it came time to capitalize the characters making up the name "Yamada", the second bytes in the pairs were erroneously treated as if they were Roman characters and "capitalized" accordingly. The result was that the name "Yamada" turned into the characters meaning "corpse" and "field". You can imagine how Mr. Yamada felt about this.

Converting the string to Unicode would have helped a little, since the Unicode capitalization rules would certainly not have connected two unrelated characters in that way. But there are still risks in character-by-character capitalization: In some languages, capitalization is itself context-sensitive. MSDN gives as an example that in Hungarian, "SC" and "Sc" are not the same thing when compared case-insensitively.

Raymond Chen

Follow